3E

УДК 004.89, 004.93

Т.В. Ермоленко, А.В. Жук
Институт проблем искусственного интеллекта МОН Украины и НАН Украины, г. Донецк
Naturewild71@gmail.com, juk@iai.dn.ua

Классификация фреймов речевого сигнала в задачах дикторонезависимого распознавания речи

В статье предлагается метод определения границ речи в поступившем речевом потоке с использованием автоматической настройки под шум окружающей среды и звукозаписывающего оборудования, а также алгоритм классификации фреймов речевого сигнала в терминах обобщенной фонетической транскрипции. Используемые параметры базируются на различных спектральных представлениях сигнала, отражают особенности спектральной плотности звуков речи, принадлежащих разным фонетическим классам, что обеспечивает дикторонезависимость процесса классификации.

Введение

Организация интерактивного взаимодействия пользователя и персонального компьютера (ПК) невозможна без средств ввода информации. Естественным способом передачи текста и команд для человека является речь. Особенно незаменим такой способ ввода для людей с ограниченными возможностями, что делает системы распознавания речи (СРР) наиболее перспективным подходом к вводу информации в ПК. Технологии распознавания речи могут стать неотъемлемой частью:

1) автоматизированных информационно-справочных систем в сетях сотовой и фиксированной связи;

2) систем госбезопасности в качестве подсистемы поиска набора ключевых слов или фраз в речевом потоке;

3) систем поиска и составления подборок записей по набору ключевых слов или фраз, предназначенных для медиа-компаний, ведущих большие базы аудио-видео данных;

4) модулей автоматического перевода в аудио- видеоаппаратуре, позволяющих просматривать фильмы на иностранных языках;

5) средств автоматической диктовки с тесной интеграцией с операционными системами и офисными приложениями, предназначенными для заполнения форм на компьютере, голосового набора текстовых сообщений на мобильном телефоне, создания электронных писем без помощи клавиатуры.

Несмотря на широкое применение технологий автоматического распознавания речи, множество проблем все еще остаются нерешенными. Высокую точность распознавания (95 – 99%) имеют командные системы, работающие с изолированными словами и малыми словарями, при этом эффективность их работы сильно зависит от уровня шума [1]. Задача распознавания слитной речи еще далека от решения, хотя именно такой тип речевого взаимодействия считается наиболее перспективным. Разработанные на сегодняшний день СРР, имеющие развитые возможности, высокую точность

и сравнительно низкие вычислительные затраты, работают с очень ограниченным словарем, требуют выполнения сложной и длительной процедуры обучения на конкретного диктора, что обусловливает невозможность их работы с неограниченным количеством постоянно сменяющихся пользователей и препятствует их широкому распространению.

Системы дикторонезависимого распознавания речи обладают большими возможностями применения и, соответственно, большей сложностью реализации. Основными проблемами, с которыми сталкиваются разработчики подобных систем, являются:

1) отсутствие методов выделения в речевом сигнале дикторонезависимых признаков;

2) недостаточная робастность алгоритмов распознавания речи к различным возможным искажениям сигнала на входе системы, вызванных шумом окружающей среды и звукозаписывающего оборудования, что приводит к значительному понижению точности работы.

Важнейшим этапом обработки речи в процессе распознавания является выделение информативных признаков, однозначно характеризующих речевой сигнал. Существует некоторое число математических методов, анализирующих речевой спектр. Здесь самым широко используемым является преобразование Фурье, известное из теории цифровой обработки сигналов [2]. Данный математический аппарат хорошо себя зарекомендовал в данной области, имеется множество методик обработки сигналов, использующих в своей основе преобразование Фурье. Несмотря на это, постоянно ведутся работы по поиску иных путей параметризации речи. Одним из таких новых перспективных направлений является вейвлет-анализ, который стал применяться для исследования речевых сигналов сравнительно недавно [3].

Робастные дикторонезависимые параметры, описывающие акустические характеристики фонетических классов звуков речи, о которых пойдет речь в данной работе, используют преобразование Фурье и вейвлет-анализ.

Цель данной работы – разработка алгоритмов сегментации и классификации РС в системах дикторонезависимого распознавания изолированных команд.

Для достижения поставленной цели необходимо решить следующие задачи:

1. Определение границ речи в поступившем речевом потоке с использованием автоматической настройки под шум окружающей среды и звукозаписывающего оборудования.

2. Классификация фреймов РС в терминах обобщенной фонетической транскрипции на базе параметров, отражающих особенности спектральной плотности звуков речи, принадлежащих разным фонетическим классам, что обеспечит дикторонезависимость процесса классификации.

Определение границ речи в звуковом сигнале

В данной работе для определения границ речи использовался аппарат вейвлет-преобразований, в частности, быстрое вейвлет-преобразование (БВП) Добеши. Вейвлет-спектр дает усредненную величину обычного спектра Фурье в окрестности центральной частоты вейвлет-фильтра, и усреднение тем грубее, чем выше частота. Таким образом, производится сглаживание спектра, что в системах распознавания речи используется для понижения чувствительности к шумам и искажениям сигнала.

Вейвлет-анализ сигналов на основе БВП аналогичен двухканальной фильтрации с помощью фильтра низких частот и полосовых фильтров с расширяющейся полосой.

В результате применения БВП к сигналу с частотой дискретизации f_d частотный диапазон разбивается на полосы фильтром низких частот с частотой среза $f_d / 2^{j max}$ и полосовыми фильтрами с полосами пропускания $[f_d / 2^{j+1}; f_d / 2^j]$, $j = \overline{1, j_{max}}$.

В работе предлагается двухэтапный VAD-алгоритм (VAD – Voice Activity Detector, детектор активности речи), который наряду с адаптацией к шуму учитывает акустические особенности широких фонетических классов звуков речи. В основе алгоритма лежит предположение о том, что первые три поступивших на вход системы буфера данных содержат только шум.

На первом этапе (обучение шуму) выполняется вейвлет-разложение сигнала $\varepsilon(n)$, содержащего образец шума, по уровням $j = \overline{1, j_{max}}$, затем этот сигнал разбивается на неперекрывающиеся фреймы. Длина фрейма ΔN зависит от периода основного тона и составляет примерно 0,02 с. Для каждого k-го фрейма сигнала $\varepsilon(n)$ по всем уровням разложения вычисляется $E_k(j)$ – энергия вейвлет-спектра сигнала. На основе массива этих значений определяются пороги:

$$\alpha(j,n) = M(E(j)) + n\sqrt{D(E(j))}, \tag{1}$$

где $M(E(j))$ и $D(E(j))$ – несмещенные оценки математического ожидания и дисперсии энергии вейвлет-спектра шума на уровне j. Для определения границ речи было выделено два множества масштабов:

– $M_s = \{1,..., j_s\}$ – соответствует высокочастотной части спектра, в которой сосредоточена энергия шумных глухих щелевых или смычно-щелевых звуков;

– $M_v = \{j_v,..., j_{max}\}$ – соответствует низкочастотной части спектра, в которой сосредоточена энергия вокализованных звуков.

Перед выполнением второго этапа (определения границ речи) вводятся две пороговые величины: *minPnL* и *maxPsL* – число фреймов, соответствующее минимальной длине фонемы и максимальной длине шумного глухого смычного звука.

На втором этапе для сигнала, содержащегося в поступившем буфере данных, выполняется вейвлет-преобразование, после чего для каждого k-го фрейма сигнала вычисляется энергия спектра $E_k(j)$ и проверяется выполнение условия:

$$BOOL(k,n) = \begin{cases} 1, & (\exists j_s \in M_s : E_k(j_s) > \alpha(j_s, n)) \vee (\exists j_v \in M_v : E_k(j_v) > \alpha(j_v, n)) \\ 0, & иначе \end{cases}$$

Если для текущего k-го фрейма $BOOL(k,n)$=1, то считается, что этот фрейм содержит речь. В наших исследованиях для определения границ речи $n = 3$.

Таким образом, номера отсчетов сигнала L и R, являющихся левой и правой границами слова, определяются согласно следующим условиям:

$$\exists l : \forall i : l < i < l + m\ (BOOL(l,3) = 0) \wedge (BOOL(i,3) = 1) \wedge (m > minPnL) \Rightarrow L = l\Delta N$$

$$\exists r > l : \forall i : r < i < r + m\ (BOOL(r,3) = 1) \wedge (BOOL(i,3) = 0) \wedge (m > maxPsL) \Rightarrow R = r\Delta N$$

Алгоритм определения границ речи в звуковом потоке можно представить в виде диаграммы состояний и переходов, изображенной на рис. 1.

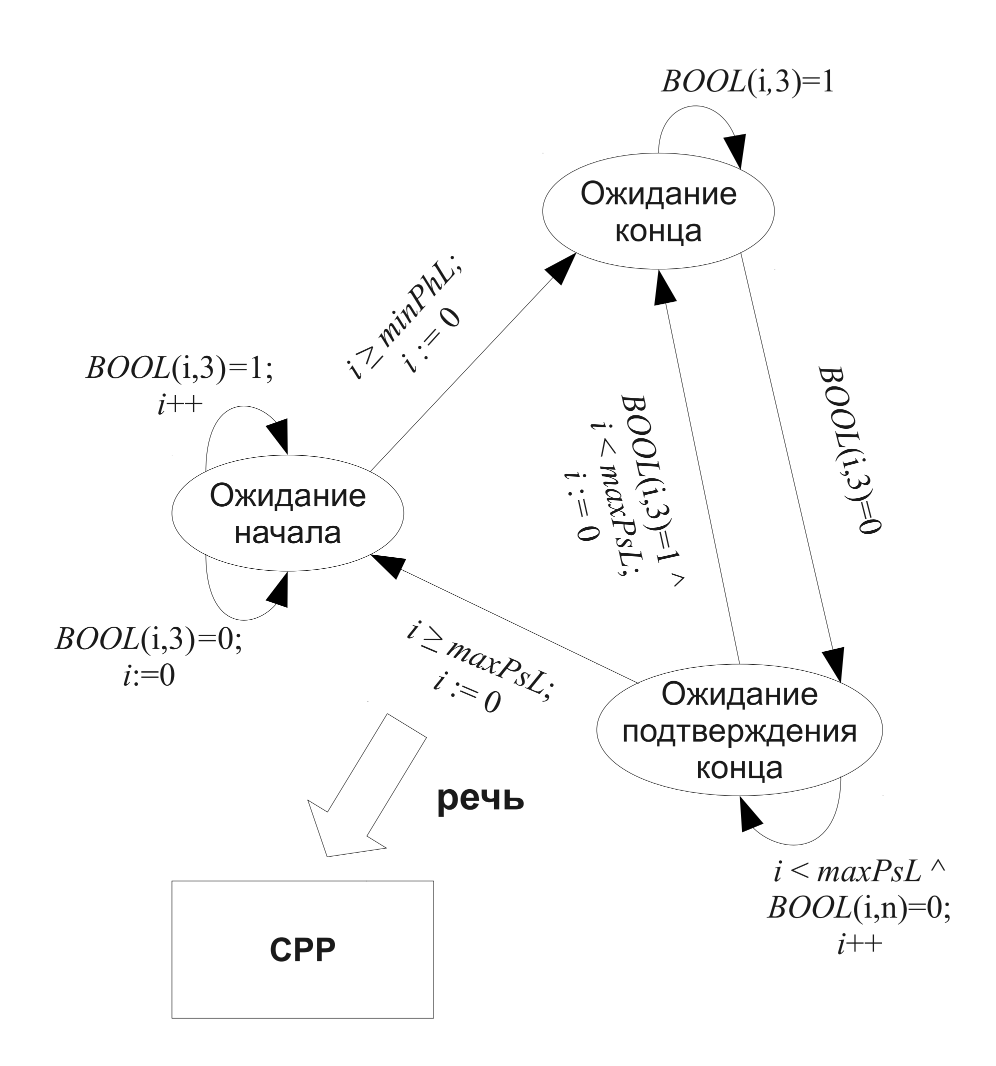


Рисунок 1 – Диаграмма состояний и переходов для определения границ речи

Классификация фреймов речевого сигнала

Звуковой сигнал, полученный в результате выполнения процедуры, описанной в п.1, может быть разбит на участки, соответствующие широким фонетическим классам (ШФК) звуков речи: вокализованным (*Voc*); шумным глухим щелевым или смычно-щелевым (*Sh*); шумным глухим смычным (*P*).

Как известно, помимо формант и основного тона, ярко выраженных в спектре вокализованных звуков, спектральная картина звуков определяется и шумовым источником – турбулентным или импульсным шумом при образовании щелевых и смычных согласных звуков, которые представлены в виде иррегулярного распределения акустической энергии во времени [4], [5]. Дикторонезависимость процедуры сегментации и классификации обеспечивают именно эти особенности спектральной плотности звуков речи, принадлежащих разным ШФК, для описания которых в работе предложено две характеристики, использующие различные спектральные представления РС, а также адаптацию под шум.

На этапе обучения шуму для каждого фрейма сигнала $\varepsilon(n)$, содержащего образец шума (первые три поступивших на вход системы буфера данных), выполняется быстрое преобразование Фурье, после чего вычисляется значение характеристики (2):

$$P(k)=\frac{\sum\limits_{j=HFBound}^{\Delta N-1}\left|FFT_j(k)\right|}{\sum\limits_{j=0}^{HFBound-1}\left|FFT_j(k)\right|},\qquad (2)$$

где k – номер фрейма; $HFBound$ – номер частоты, соответствующей левой границе высокочастотной части спектра, в которой сосредоточена энергия звуков из класса Sh (около 4 кГц); $\{FFT_j\}_{j=0}^{\Delta N-1}$ – массив коэффициентов Фурье-спектра, полученный для k-го фрейма.

По аналогии с алгоритмом определения границ речи, для адаптации под шум на основе несмещенных оценок математического ожидания $M(P)$ и дисперсии $D(P)$, полученных по массиву значений (2) для сигнала $\varepsilon(n)$, определяется порог:

$$\alpha(n)=M(P)+n\sqrt{D(P)}\qquad (3)$$

Величина (2) характеризует отношение энергии спектра в высокочастотной области к энергии в низкочастотной. Очевидно, что для фреймов, содержащих звук, спектр которых сосредоточен в области высоких частот, значение $P(k)$ будет больше 1. К таким звукам, помимо шумных щелевых, могут относиться и вокализованные звуки, например, гласные верхнего подъёма ([и], [э]), которые всегда отличает относительно большая роль высших формант, что сказывается на поведении плотности распределения спектра [4], [5]. Для большинства вокализованных звуков спектр сосредоточен в области 1,5 кГц, следовательно, значения $P(k)$ не превысят 1. Значения (2), полученные для шумных глухих смычных (паузоподобных) звуков, меньше порога (3). Таким образом, в набор решающих правил для классификации фреймов входят следующие:

$$P(k)\geq\alpha(n)\Rightarrow k\in Sh\vee k\in Voc\,,\; P(k)<\alpha(n)\Rightarrow k\in P\vee k\in Voc\qquad (4)$$

Поведение характеристики (2) для звуков разных классов наглядно демонстрирует рис. 2, на котором показаны графики амплитудно-временного представления (АВП) и значений $P(k)$ для реализации слова «шесть», горизонтальная линия соответствует значению порога (3) при $n = 5$.

Для окончательной классификации на классы Voc, Sh и P используются значения вейвлет-спектра на множестве уровней разложения M_v. Считается, что поступивший на вход алгоритма классификации РС обязательно содержит вокализованные звуки, энергия которых на множестве уровней разложения M_v существенно больше энергий невокализованных, к которым относятся звуки классов Sh и P. Как показали исследования, значения энергий вейвлет-спектра на этих уровнях разложения для малоамплитудных шумных звонких щелевых ([в], [в’]) и шумных звонких смычных ([б], [д], [г], [б’], [д’], [г’]) согласных сравнима с энергией спектра ударных гласных, а значения энергий звуков классов Sh и P составляет менее 10% от энергии высокоамплитудных гласных. На рис. 3, 4 показан график АВП реализации слова «шесть» (рис. 3) и слова «два» (рис. 4), а также результат БВП на уровнях j=5 (верхний график) и j=6 (нижний график), горизонтальные прямые соответствуют значению, составляющему 10 % от максимального значения коэффициента вейвлет-спектра РС на соответствующих уровнях.

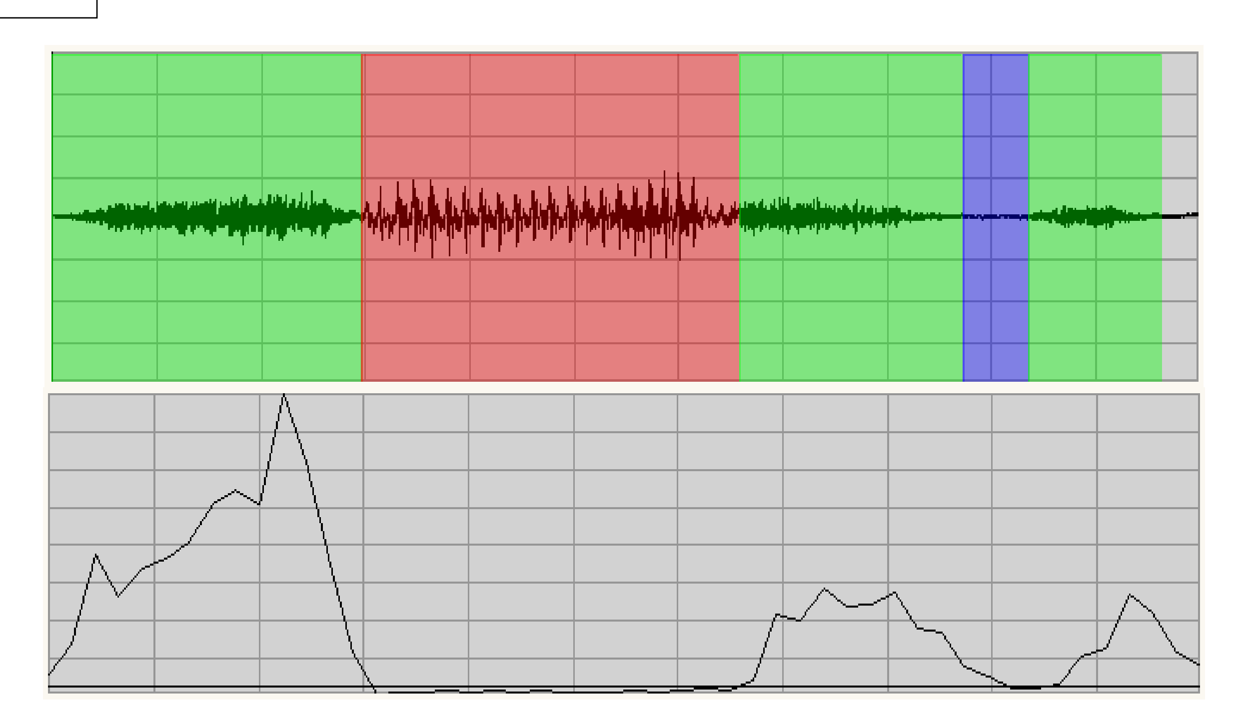


Рисунок 2 – Графики АВП (вверху) и значений *P*(*k*) (внизу), полученные для реализации слова «шесть»

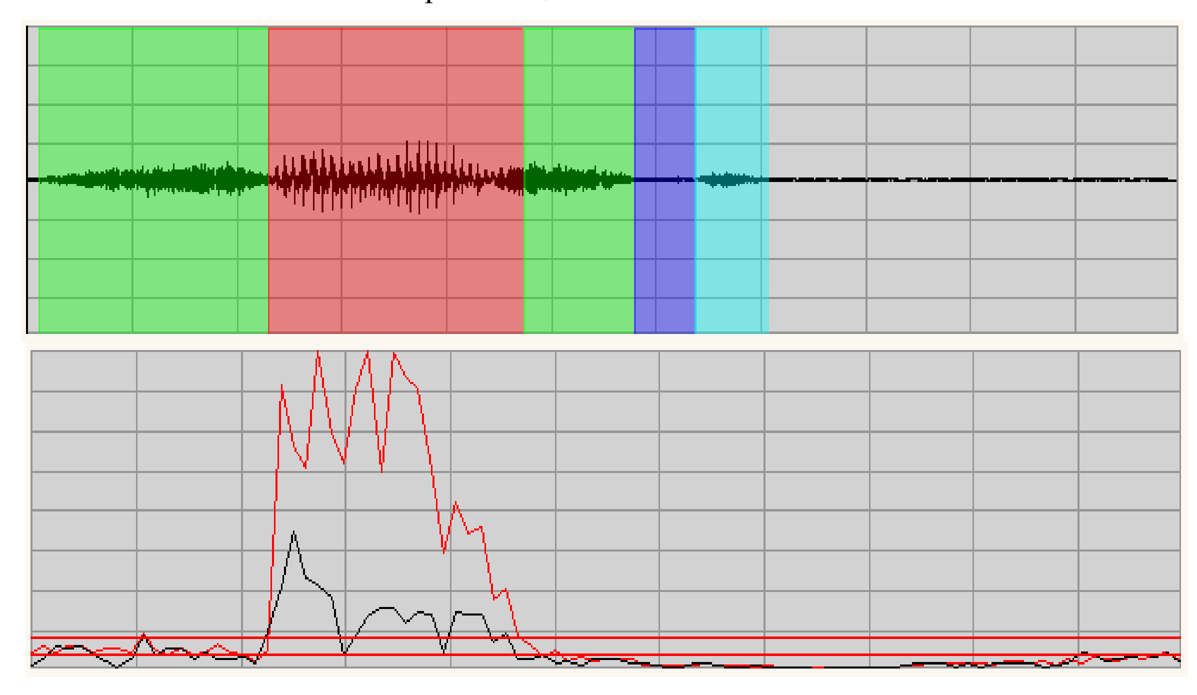


Рисунок 3 – Графики АВП (вверху) и коэффициентов вейвлет-спектра при *j*=5,6 (внизу), полученные для реализации слова «шесть»

Исходя из этих соображений для классификации используется характеристика (5) и набор решающих правил (6):

$$BoolW(k)=\begin{cases}1, \text{ если } \exists j_v \in M_v : E_k(j_v) > 0.1 maxLevEn(j_v)\\ 0, \text{ иначе,}\end{cases} \quad (5)$$

где *maxLevEn*(*j*) – максимальное значение энергии коэффициента вейвлет-спектра на уровне *j*.

$$BoolW(k)=0 \Rightarrow k \in Sh \vee k \in P\,,\ BoolW(k)=1 \Rightarrow k \in Voc \qquad (6)$$

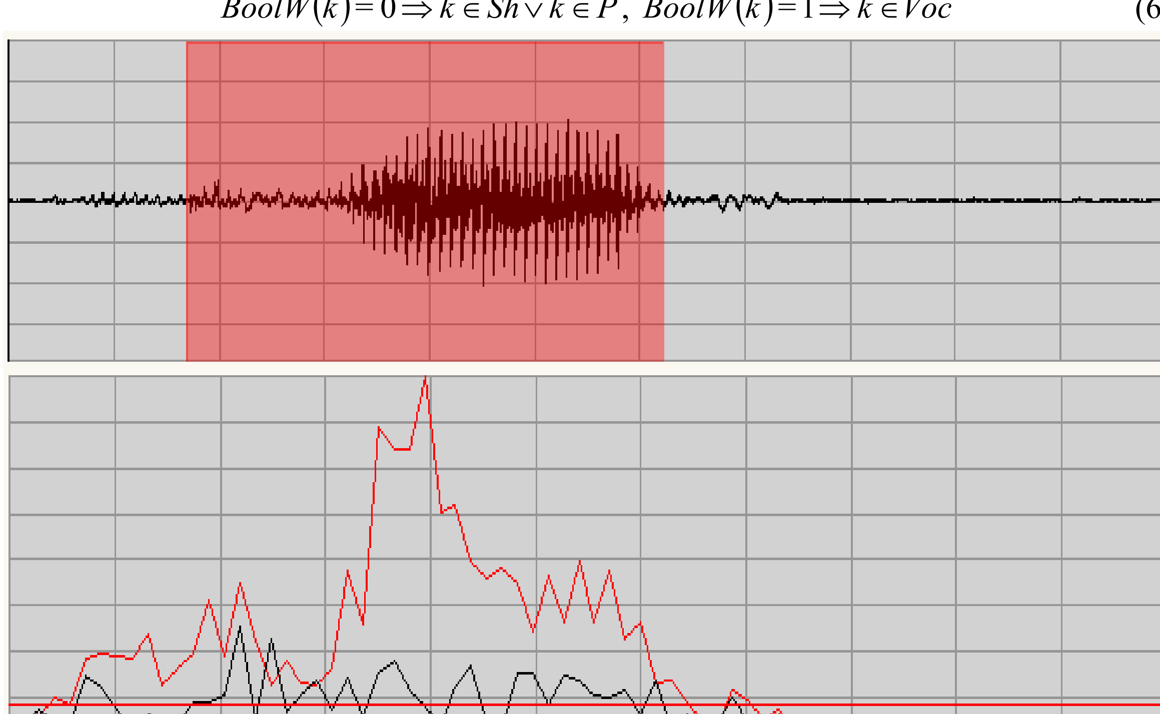


Рисунок 4 – Графики АВП (вверху) и коэффициентов вейвлет-спектра при *j*=5,6 (внизу), полученные для реализации слова «два»

Таким образом, особенности спектральной плотности звуков речи каждого из ШФК, описываемые характеристиками (2), (5), позволяют провести классификацию фреймов РС по набору правил (4), (6), как показано на рис. 5.

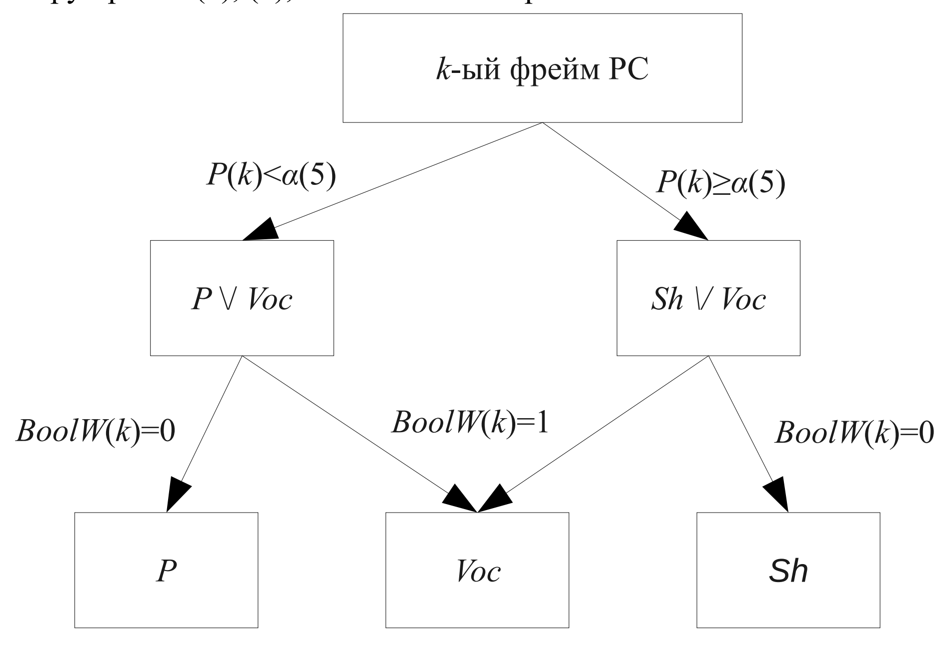


Рисунок 5 – Дерево решений для классификации фреймов РС

По результатам классификации фреймов РС легко провести сегментацию: подряд идущие фреймы, принадлежащие одному ШФК, объединяются в один сегмент. По полученной сегментации строится обобщенная фонетическая транскрипция, которая позволяет описать общую структуру слова, т.е. модель чередования гласных, согласных, шипящих и т.д. В русском языке слов с одинаковой структурой относительно мало, например, на 100-тысячный словарь Зализняка максимальное число слов с одинаковой структурой – около 150, то есть меньше 0,2 %. Таким образом, в результате обобщенной классификации выводятся в качестве сокращенного списка кандидатов на распознавание только те слова, которые имеют ту же структуру, что и распознаваемое. Верное распознавание последовательности классов при любых ошибках внутри классов приводит к значительному сокращению числа слов-кандидатов на распознавание.

Исследование эффективности методов

Было проведено численное исследование предложенных алгоритмов на 10 дикторах (мужчин и женщин с разными голосовыми данными), каждый из которых произносил по 15 слов, содержащих фонемы разных классов. Запись проводилась в формате WAV PCM с частотой дискретизации 22 050 Гц и глубиной квантования 16 бит с помощью микрофонов с разными характеристиками. Для вычисления порога (1) использовался параметр n=3, для вычисления порога (3) – n=3. БВП выполнялось при $j_{max} = 6$, полученные вейвлет-спектры анализировались на уровнях $M_s = \{1,2\}$ и $M_v = \{5,6\}$.

Качество классификации сегментов оценивалось по функционалу:

$$F = \frac{m}{n} \to max, \qquad (7)$$

где n – общее количество сегментов, принадлежащих разным ШФК, по всем реализациям слов всех дикторов, m – количество правильно классифицируемых сегментов.

Вероятность правильной классификации сегментов, определенная по формуле (7), составила 0,984. Полученное высокое значение вероятности правильной классификации на речевом материале, принадлежащим разным дикторам, свидетельствует об эффективности разработанных алгоритмов и перспективности предложенного подхода применительно к задачам дикторонезависимого распознавания речи.

Выводы

Рассмотренный в работе подход позволяет достаточно надежно и дикторонезависимо выделять в звуковом потоке фрагменты, содержащие речь, и выполнять сегментацию этих фрагментов с одновременной классификацией сегментов по широким фонетическим классам. Эффективность предложенного подхода подтверждается проведенными исследованиями, в результате которых вероятность правильной классификации сегментов составила 0,984. Дальнейшие исследования в данном направлении предусматривают поиск робастных признаков как для более детальной классификации сегментов, так и для распознавания слова как единого целого.

Литература

1. Леонович А.А. «Современные технологии распознавания речи» [Электронный ресурс]. – Режим доступа: http://masters.donntu.edu.ua/2008/kita/kravchenko/library/artone.htm

2. Рабинер Л.Р. Цифровая обработка речевых сигналов / Л.Р. Рабинер, Р.В. Шафер ; пер. с англ. – М. : Радио и связь, 1981. – 496с.
3. Малла С. Вейвлеты в обработке сигналов / Малла С. ; пер. с англ. – М. : Мир, 2005. – 671с.
4. Фант Г. Акустическая теория речеобразования / Фант Г. ; пер. с англ. – М. : Наука, 1964. – 326с.
5. Златоустова Л.В. Фонетические единицы русской речи / Златоустова Л.В. – М. : МГУ, 1981. –108 с.

Literatura

1. Leonovich A.A. «Sovremennye tekhnologii raspoznavaniya rechi» [Electronic resource] Access mode: http://masters.donntu.edu.ua/2008/kita/kravchenko/library/artone.htm
2. Rabiner L.R., Shafer R.V. Tsifrovaya obrabotka rechevykh signalov: Per. s angl. M.: Radio i svyaz', 1981. 496s.
3. Malla S. Veyvlety v obrabotke signalov: Per. s angl. M.: Mir, 2005. 671s.
4. Fant G. Akusticheskaya teoriya recheobrazovaniya: Per. s angl. M.: Nauka, 1964. 326s.
5. Zlatoustova L.V. Foneticheskie edinitsy russkoy rechi. M.: MGU, 1981. 108 s.

Т.В. Єрмоленко, О.В. Жук

Класифікація фреймів мовленнєвого сигналу в задачах дикторонезалежного розпізнавання мовлення

У статті запропоновано метод визначення границь мовлення у потоці мовлення, що надійшов на вхід системи розпізнавання, з використанням автоматичного налаштування під шум оточуючого середовища та звукозаписуючого обладнання, а також алгоритм класифікації фреймів мовленнєвого сигналу у термінах узагальненої фонетичної транскрипції. Параметри, що було використано, базуються на різних спектральних представленнях сигналу, відображають особливості спектральної щільності звуків мовлення, які належать до різних фонетичних класів, що забезпечує дикторонезалежність процесу класифікації.

T.V. Yermolenko, A.V. Zhuk

Speech signal frames classification in the tasks of speaker-independent speech recognition

The method for voice activity detection in a captured speech stream with automatic adaptation to environmental and sound-capture hardware noises, and the algorithm for speech signal frames classification in the terms of generalized phonetic transcription are proposed in the article. The speaker-independence is reached because of the parameters used in the classification process. These parameters are based on different spectral representations of a signal and reflect spectral density species of speech sounds.

Статья поступила в редакцию 01.06.2011.